Assessing the Predictive Influence of Human, Road Conditions, and Vehicle Factors on Accident Severity in Victoria

W13 G7 – Tazeen Atif, Nikhil Gaba, Shwethan Potu

RESEARCH QUESTION & OBJECTIVES OF ANALYSIS

- To what extent are **predetermined factors**, including seat belt usage, age, road surface conditions, atmospheric conditions, node type, fuel and vehicle types, **predictive** of **accident severity** on Victorian roads?

- The primary objective of the analysis is to **investigate factors that contribute to road accident severity** in Victoria, by leveraging data preprocessing, correlation analysis, and SLM models. This helps identify the most impactful risk indicators, enabling data-driven strategies to improve road safety and injury prevention

SUMMARY OF DATASETS

Filtered Vehicle

A smaller csv file of the vehicle dataset (from A1) that includes key characteristics like **vehicle type and fuel type** for each vehicle in an accident.

Road Surface Condition

Contains information about road surface conditions at the time of each accident. It includes **encoded values and descriptive labels** (dry, wet, icy) for accidents.

Atmospheric Condition

Provides data on atmospheric conditions during accidents (clear, rain, fog), and also uses **encoded values and descriptive labels.**

Node

Describes the geographic details of each accident location, especially **node type (like intersection).**

Person

Contains information about every individual involved in the accidents, including **age groups, injury level, and seatbelt or helmet use.**

Accident

The core dataset summarises each accident. Our investigation had its target variable from this dataset – **accident severity.**

PREPROCESSING TECHNIQUES

One-hot Encoding

- Convert categorical data to **numerical** format for further analysis (correlation and supervised learning models)
- Label Encoding was not chosen because:
  - It introduces notions of **distance** between categories & Ordinal Relationships

Discretisation

- Convert Age, continuous numerical data, into discrete numerical format
- Ensures consistency between features, and allows one correlation measure to be used

Merging Datasets

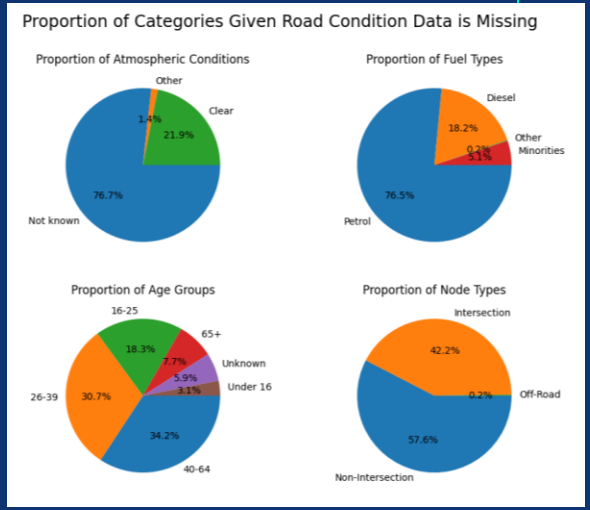
- Merged relevant features from each of the chosen datasets, with accident severity into one data object

PREPROCESSING – WHY IMPUTATION?

	AGE GROUP	HELMET BELT WORN	ROAD SURFACE	ATMOSPHERIC CONDITION	NODE TYPE	FUEL TYPE	VEHICLE TYPE
PROPORTION UNKNOWN VALUES	3.17%	27.08%	6.35%	9.97%	<0.01%	2.44%	0%

Before analysing, we handled missing values

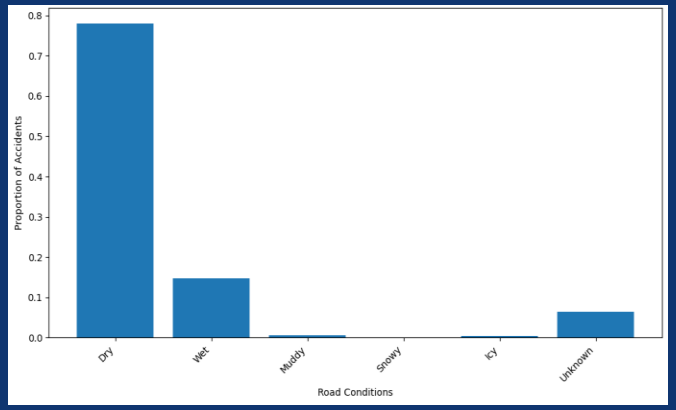
- Most missing values were either MCAR or MAR, so were handled via **mode imputation**
- EXCEPT where doing so introduced bias – 'Helmet Belt Worn' had 27% missing
- OR where it was unnecessary - 'Vehicle Type' has 0% missing



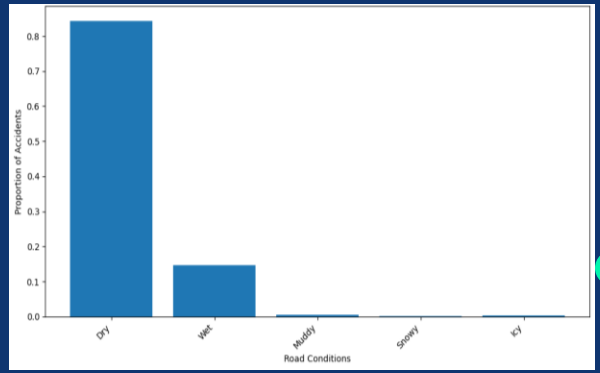
DISTRIBUTION BAR CHARTS

ROAD SURFACES

BEFORE IMPUTATION



AFTER IMPUTATION

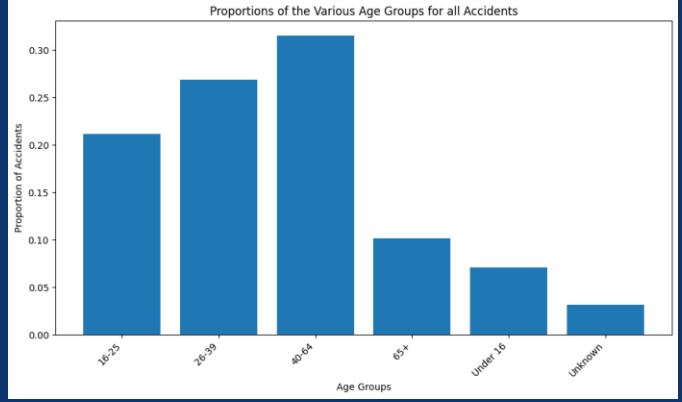


- Low proportion of ‘unknown’ values imputed as the most frequent category, appearing almost 80% of the data
- Distribution of the categories before and after imputation is quite similar, suggesting low bias

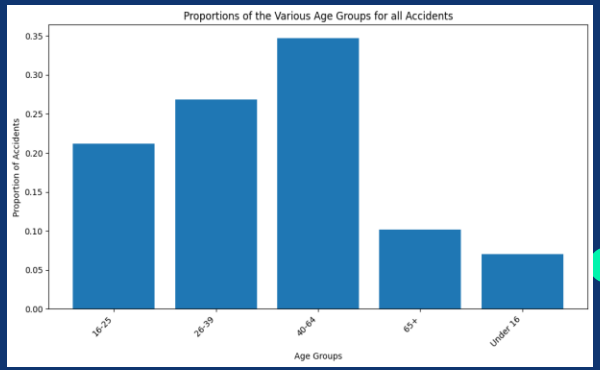
DISTRIBUTION BAR CHARTS

AGE GROUP

BEFORE IMPUTATION



AFTER IMPUTATION

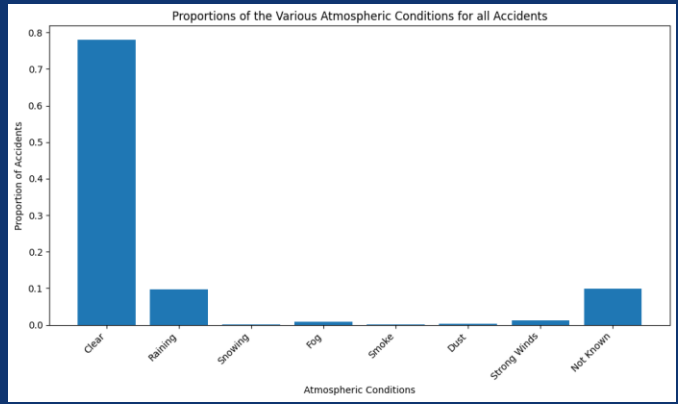


- Low proportion of ‘unknown’ values imputed as the most frequent category
- Distribution of the categories before and after imputation is quite similar, suggesting low bias

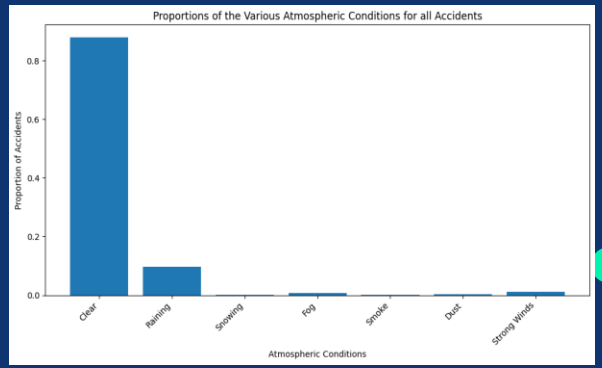
DISTRIBUTION BAR CHARTS

ATMOSPHERIC CONDITIONS

BEFORE IMPUTATION



AFTER IMPUTATION

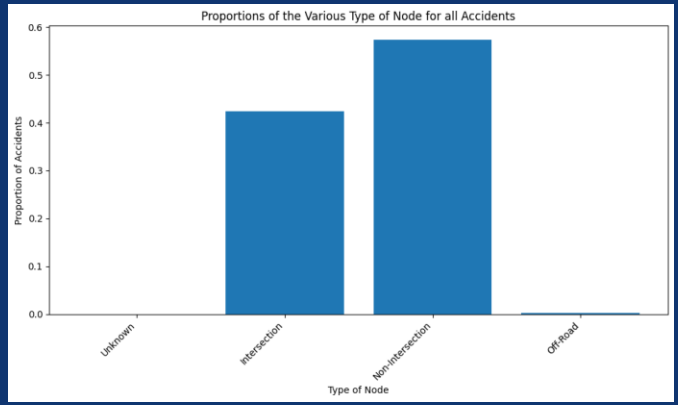


- Low proportion of ‘unknown’ values imputed as the most frequent category, appearing almost 80% of the data
- Distribution of the categories before and after imputation is quite similar, suggesting low bias

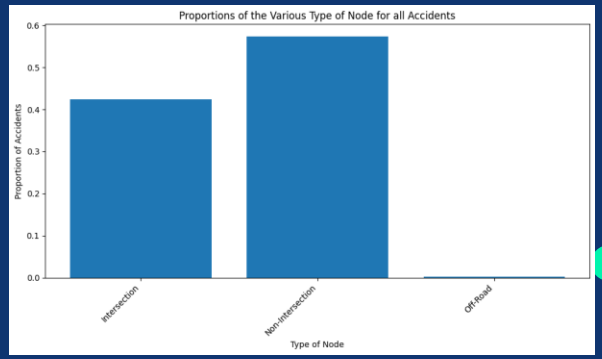
DISTRIBUTION BAR CHARTS

NODE TYPE

BEFORE IMPUTATION



AFTER IMPUTATION

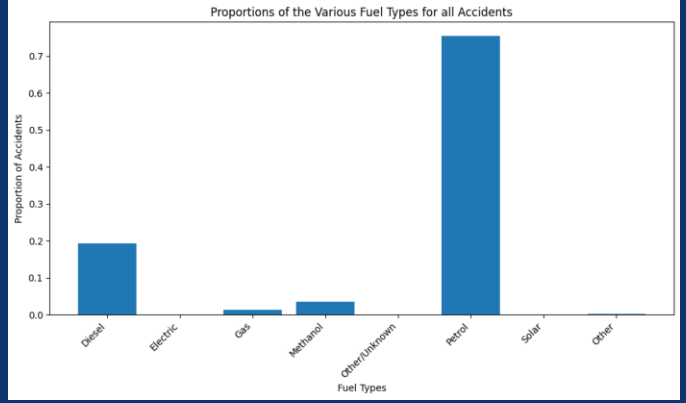


- Low proportion of ‘unknown’ values imputed as the most frequent category
- Distribution of the categories before and after imputation is quite similar, suggesting low bias

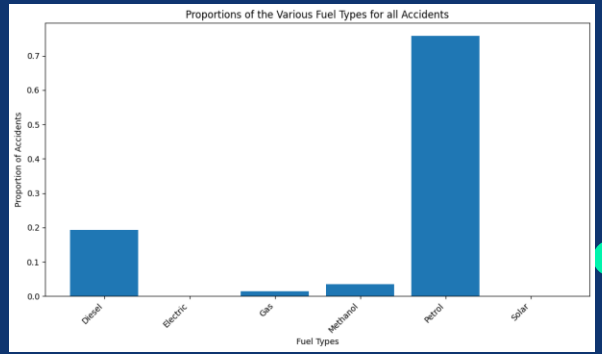
DISTRIBUTION BAR CHARTS

FUEL TYPE

BEFORE IMPUTATION



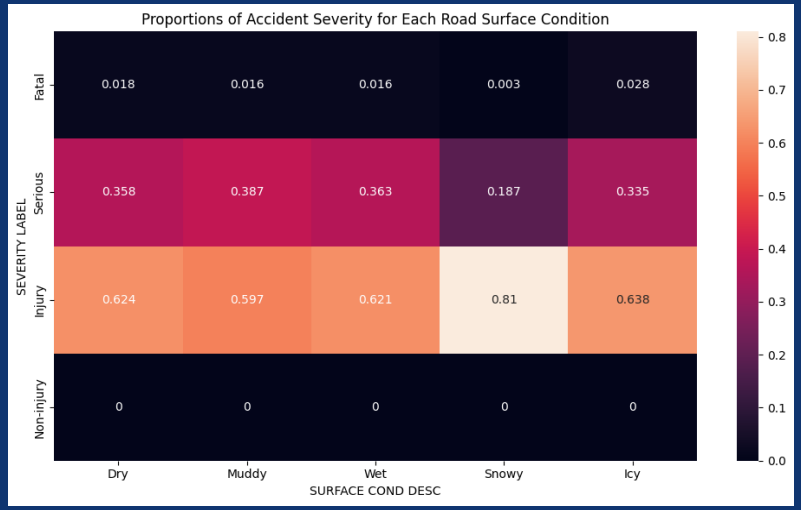
AFTER IMPUTATION



- Low proportion of ‘unknown’ values imputed as the most frequent category, appearing almost 80% of the data
- Distribution of the categories before and after imputation is quite similar, suggesting low bias

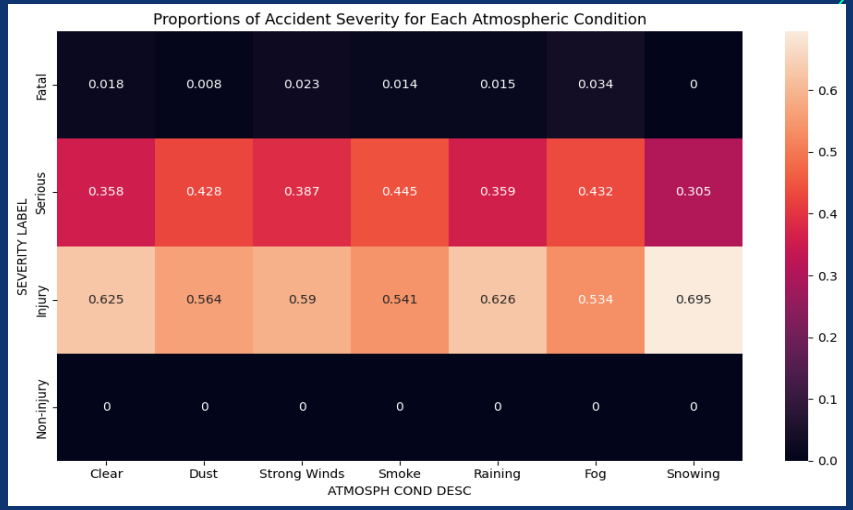
DATA VISUALISATION – PROPORTION HEATMAPS

ROAD SURFACE



- Low Proportion of Serious accidents in Snow (18.7% vs ~35% for other surfaces)
- Other Surfaces have similar proportions
- Drivers drive slower in Snowy conditions

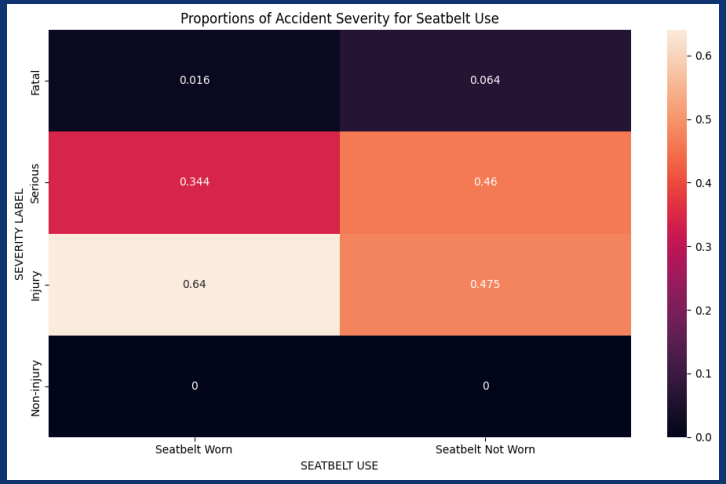
ATMOSPHERIC CONDITIONS



- Dust, Fog and Smoke have severe accidents involving higher impact (~43% Serious)
- Other Conditions (Clear, Strong Winds, Raining, etc.) are usually less severe
- Visibility could potentially explain this

DATA VISUALISATION – PROPORTION HEATMAPS

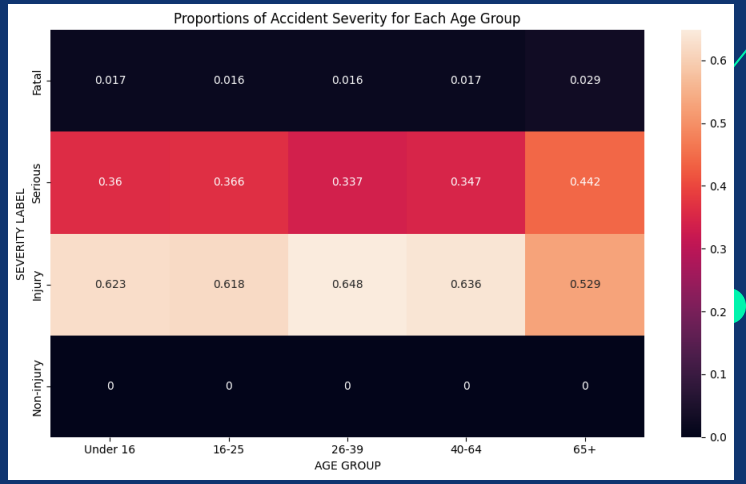
SEATBELT USAGE



Wearing a Seatbelt is generally associated with reduced accident severity:

- 46% of Serious Accidents Not Wearing Seatbelt vs 34.4% Wearing a Seatbelt
- 6.4% Fatality Rate Not Wearing Seat Belt vs 1.6% when Wearing a Seat Belt

AGE GROUP

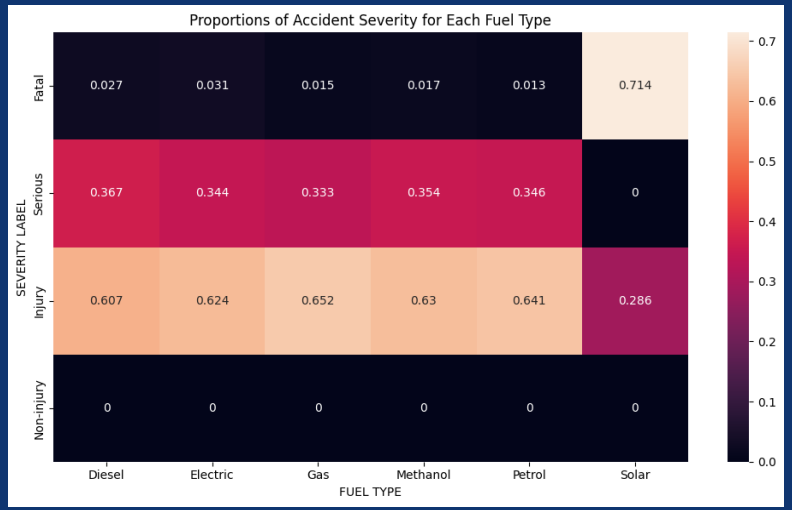


Accidents involving the 65+ Age Group tend to be more severe in terms of injury:

- 44.2% Serious vs ~35% for Younger Ages
- 2.9% Fatal vs ~1.6% for Younger Ages
- 52.9% Non-Serious vs ~63% for Younger Ages

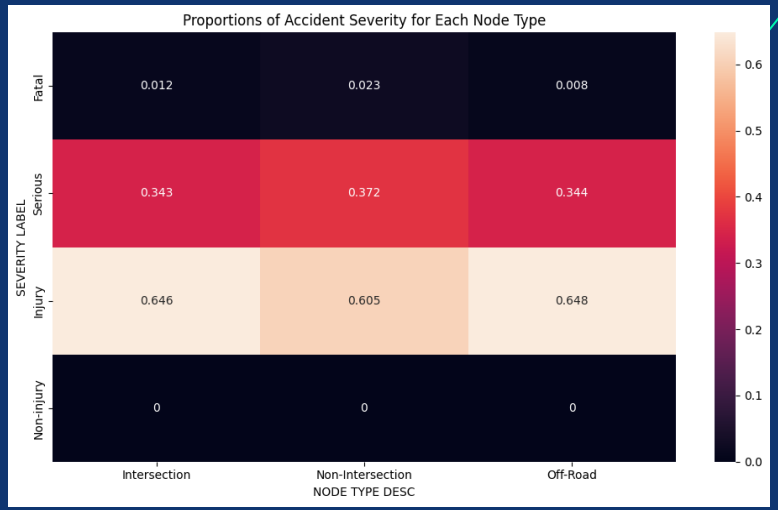
DATA VISUALISATION – PROPORTION HEATMAPS

FUEL TYPE



- Solar-powered vehicles show an unusually high fatality rate, likely due to a very small sample size
- Conventional fuel types show consistent patterns resulting in injury
- Serious injuries account for about one-third of cases across all mainstream fuel types

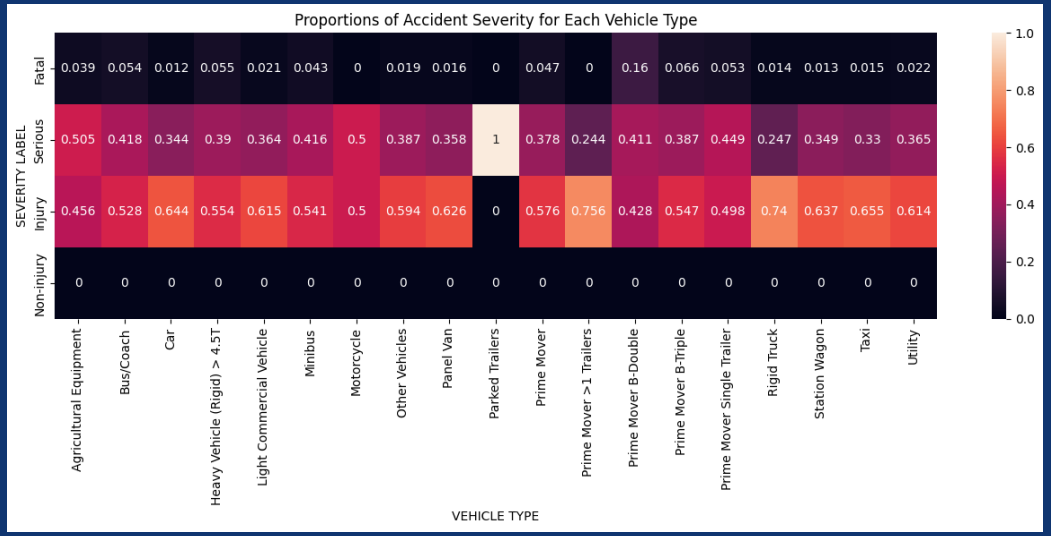
NODE TYPE



- Non-intersections (freeways, etc.) have a higher proportion of Serious Injury or Fatal Accidents, potentially because of higher speeds on roads
- Intersections and Accidents Off-Road are usually less severe

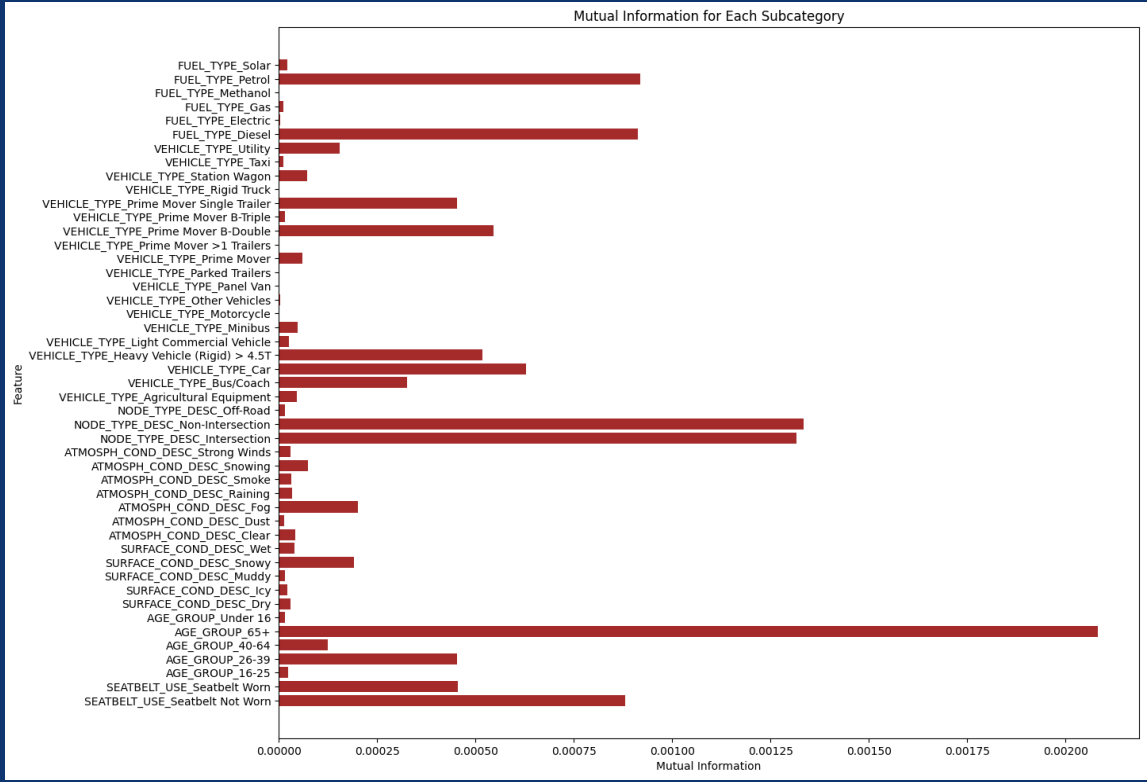
DATA VISUALISATION – PROPORTION HEATMAPS

VEHICLE TYPE



- Heavier commercial vehicles are associated with higher proportions of injuries (75.6%) and fatalities (16%), indicating more severe accident outcomes
- Cars and light commercial vehicles show more moderate proportions of injuries ranging from 60—65% and lower fatality rates of 1-2%
- Parked trailers show 100% serious injury rate, which likely reflects a very small and skewed sample size, so this should be interpreted with caution

DATA VISUALISATION – MUTUAL INFORMATION



- Mutual Information (MI) scores were computed to assess feature relevance
- It measured non-linear dependencies between each predetermined factor and the target variable – accident severity
- **'Node Type Intersection,' 'Node Type Non-Intersection'** and **'Age Group 65+'** had the highest, relative MI scores
- MI correlation guided feature selection for our classification models

DECISION TREE – CLASSIFICATION REPORT

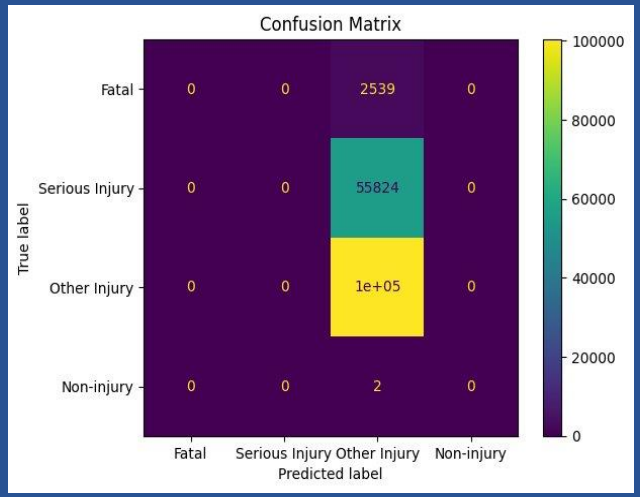
	PRECISION	RECALL	F1-SCORE	SUPPORT
FATAL	0.000	0.000	0.000	2539
SERIOUS	0.000	0.000	0.000	55842
OTHER	0.632	1.000	0.775	100360
NON-INJURY	0.000	0.000	0.000	2
ACCURACY			0.632	158752
MACROS AVG	0.158	0.250	0.194	158752
WEIGHTED AVG	0.400	0.632	0.490	158752

K-NN – CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
FATAL	0.000	0.000	0.000	2539
SERIOUS	0.418	0.058	0.102	55842
OTHER	0.636	0.956	0.764	100360
NON-INJURY	0.000	0.000	0.000	2
ACCURACY			0.625	158752
MACROS AVG	0.263	0.254	0.216	158752
WEIGHTED AVG	0.549	0.625	0.519	158752

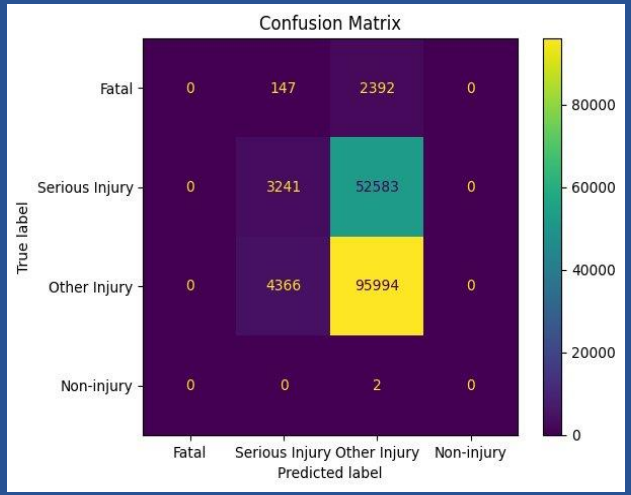
CONFUSION MATRICES

DECISION TREE



Confusion matrix shows all 'Serious' and 'Fatal' accidents were misclassified - indicating severe class imbalance issues and overfitting to the majority class

K-NN



The K-NN confusion matrix shows that the model heavily favoured the majority class ‘Other Injury’. The matrix shows majority of ‘Fatal’ and ‘Serious Injury’ were misclassified, indicating sever class imbalance

CONCLUSION

To what extent are predetermined factors, including seat belt usage, age, road surface conditions, atmospheric conditions, node type, fuel and vehicle types, predictive of accident severity on Victorian roads?

Our analysis found that while factors like seatbelt use, elderly age, and heavy vehicle types show some correlation with accident severity, their **predictive power is limited**. Classification models **struggles with class imbalance**, often defaulting to the majority class. Still, these findings **highlight key risk areas** for targeted road safety measures, and **enables decisions** such as safety campaigns towards seatbelt use, and regulatory measures for heavy vehicles to be made which further reinforces the value of data-driven decisions